A Machine Learning Approach to Customer Churn Prediction in Telecom Services

Chinmaya Gouda
School of Computer Science Engineering
Lovely Professional University Phagwara, India
Chinmayakumaar8888@gmail.com

Abstract:

Customer churn remains a challenge for telecommunications companies, since a customer churn can be very costly to replace. Traditional methods in customer churn analysis appear not to be effective in identifying complicated customer behavior due to their dependency on manual rules and simple statistics. The research will employ machine learning algorithms to model customer churn based on demographic, usage, and billing variables.

The Telco Customer Churn dataset contains variables such as tenure, type of contract, payment method, internet service, and monthly charges. A variety of supervised learning classifiers such as Logistic Regression, Decision Trees, Random Forest, and XGBoost are tested and validated. The performance of these models is measured in terms of accuracy, precision, recall, F1-score, and ROC AUC curve. Experiments conducted show that Logistic Regression performed better on this particular dataset, which performed better than other complicated ensemble methods. Moreover, an easy-to-interact dashboard is presented to illustrate the trends and results of this model.

Keywords—
Customer Churn, Machine Learning, Telecom Analytics, Random Forest, XGBoost, Classification

I. INTRODUCTION

Customer churn is the term used to describe when clients stop using a business's services. Churn has a direct impact on revenue and long-term business viability in the fiercely competitive telecom sector. Large amounts of consumer data about contracts, billing, and service usage are gathered by telecom companies. Nevertheless, it is still difficult to draw useful conclusions from this data using conventional analysis techniques.

Traditional churn prediction methods frequently depend on simple statistical models or rule-based systems. Nonlinear relationships between customer attributes and churn behaviour are difficult for these methods to capture. The capacity of machine learning techniques to extract intricate patterns from massive datasets has drawn attention in recent years.

Using the Telco Customer Churn dataset, this project offers a machine learning-based framework for customer churn prediction. To determine the best strategy, a number of classification models are put into practice and contrasted. Additionally, an interactive dashboard that bridges the gap between analytical models and real-world business use is created to present churn insights in an approachable way.

II. RELATED WORK

Customer churn prediction is a widely explored task in both data mining and machine learning studies. Initial research work used both logistic regression and decision trees for handling customer churn. Although these approaches were interpretable and provided understandable results, they were not accurate in handling complex datasets.

Later studies focused on learning with ensembles such as Random Forest and Gradient Boosting Machines, which indicated higher accuracy levels. Methods such as these handle both mixed and nonlinear relationships of variables in customer behaviour very well.

Recent studies have underlined the efficiency of the gradient boosting techniques on XGBoost for performing churn prediction. These models, while outperforming other traditional classifiers, focus on misclassified instances and optimize performance iteratively. However, many existing studies focus only on model accuracy and lack deployment-oriented visualization or interpretability for business users.

The paper extends prior work by benchmarking various machine learning models on a real-world telecom dataset and integrates the results in an interactive dashboard for better interpretability and practical use.

III. METHODOLOGY

The proposed system follows a structured machine learning pipeline consisting of data preprocessing, exploratory data analysis, feature engineering, model training, evaluation, and visualization.

A. Data Description

The dataset being used in this research is Telco Customer Churn, which contains a wide array of information with regards to customers of a telecommunications service provider. The dataset comprises a total of 7,043 entries or rows and a total of 21 variables or columns, with each column symbolizing a unique customer.

Every record contains a mix of demographic, service subscription, and billing-oriented information, making this dataset ideal for behaviour analysis and churn modelling. The dependent variable, Churn, representing a customer abortion of service, is used.

Dataset Composition

- **Number of records:** 7,043
- **Number of features:** 21
- **Target variable:** Churn (Yes / No)
- **Problem type:** Supervised binary classification

B. Data Preprocessing

Preprocessing was done to prepare the dataset in several steps:

1. Handling Missing Values:

There were a lot of missing values in the TotalCharges feature because of inappropriate data type representation. The missing values were numerically changed, and rows containing those were removed.

2. Removal of Non-informative Features:

The customerID column was removed, as it does not contribute to the churn prediction process

3. Categorical Encoding:

Categorical variables such as contract type, payment method, and internet service were transformed using one-hot encoding to make them suitable for machine learning models.

4. Feature Scaling:

For development status, we turned it into a binary feature. Producing or past-producing sites became 1, occurrences became 0. This matters because status shows confidence in geology and economic worth. Producing spots often have conditions good for mineralization, so it helps predict gold potential.

C. Feature Engineering

Feature engineering focused on improving predictive capability:

- Conversion of churn labels to binary format
- One-hot encoding of categorical attributes
- Retention of key numerical features such as tenure and charges

These steps resulted in a fully numerical dataset suitable for model training.

IV. MODEL SELECTION

To ensure a fair and comprehensive comparison, multiple supervised classification models were selected and implemented. These models represent a progression from simple, interpretable techniques to advanced ensemble-based methods. Using a diverse set of models helps evaluate how increasing model complexity affects churn prediction performance.

1. Logistic Regression:

Logistic Regression is chosen as a baseline model because of its simplicity and ease of understanding. Logistic Regression attempts to represent a linear combination of input variables with a function called a logistic function to express the chance of a customer leaving. Logistic Regression gives a deep insight into each input variable such as tenure or monthly charges and how these impact a customer leaving. Yet, in circumstances where complicated interplay among variables is common in customer behaviour data, Logistic Regression can be a problematic algorithm.

2. Decision Tree Classifier:

The Decision Tree classification is the appropriate choice because it allows the visualization of how various characteristics of customers are connected regarding whether or not they will stay or leave the company. By continuously dividing the initial dataset based on feature attributes, we create a decision-tree type structure that indicates how to make further classifications. Also, since decision trees are easy to visualize, they provide more assistance in understanding the reasons customers may choose to leave. The biggest downside of a single decision tree is that it can

misclassify data (known as overfitting) when it gets too complicated and starts focusing on irrelevant data patterns.

3. Random Forest Classifier:

The Random Forest technique is an ensemble modelling technique made up of many individual trees created as a collection of trees using different subsets of the original dataset. When multiple trees combine their predictions, Random Forest reduces variance and improves generalizability of the results when compared to one single tree. Random Forest will work well with mixed numerical and categorical datasets like those in Telecom customer data. Random Forest will be used to examine advantages of ensemble learning and provide improved stability for predictive capabilities.

4. XGBoost Classifier:

XGBoost (Extreme Gradient Boosting) is an advanced version of the Ensemble Learning Model with a strong emphasis on obtaining very high levels of predictive accuracy. As opposed to Random Forests which create their individual trees independently, the trees in XGBoost are created one after another and are intended to correct any previous errors made during the earlier stages. Therefore, it's ability to find nonlinearities, complex interactions among multiple features and many other very fine distinctions when identifying customers that may leave is a result of the models' usage of a boosting method. XGBoost has also built-in regularization considerations that limit overfitting and creates stability or strength when evaluating structured datasets. XGBoost's numerous strengths have led to superior predictive performance when predicting which customers are likely to "leave".

V. TRAINING AND VALIDATING STRATEGY

Stratified sampling was used to maintain the original distribution of churn classes for the training/testing sets, which were split 80:20. The training data for all models consisted of the same set of examples, and test sets were composed of a single unseen sample for each model. By using the same training and testing data, an accurate and unbiased comparison of the performance of each model could be made. Consequently, performance assessments were made on test data to predict in the real world and minimise the possibility of overfitting.

VI. PERFORMANCE METRICS

The various evaluation metrics were employed to examine the performance of the classification models. Using a single metric (accuracy) for evaluating churn prediction problems is risky since accuracy may be misleading for imbalanced data sets.

Accuracy measures the overall percentage of correctly identified customers classified as positive and negative.

Precision measures how accurate the positive predictions were (i.e., what percentage of customers predicted to have churned actually did churn).

Recall measures how well the model was able to identify and detect churned customers. This is extremely important when predicting customer churn because if you do not identify a customer that has churned you will lose revenue.

F1-Score combines precision and recall into one number to give an overall balanced measure of the model's performance.

ROC–AUC evaluates how well the classification model distinguishes between churned customers and non-churned customers at various classifications.

By using a combination of these metrics together, one obtains a comprehensive overview of the model's overall performance.

VII. RESULTS AND DISCUSSION

To assess all four models for classification, Accuracy, Precision, Recall, F1 Score, and ROC AUC were used. The results have shown a varying performance of all four models on different metrics, which signifies the significance of using all metrics to assess chum evaluation.

Model Performance Comparison

Logistic Regression scored the highest accuracy of 0.804 and an ROC-AUC of 0.836 among all models. Additionally, it scored the highest F1-score of 0.609, which implies a balance in terms of both precision and recall. Logistic Regression performed well despite making linear assumptions because it generalized well on the testing set.

The accuracy and ROC-AUC score for the Decision Tree model were lower, with a value of 0.785 and 0.815, respectively. While it did manage to learn some non-linear relationships, the lower precision and recall suggest an increased number of misclassifications, which can be a consequence of overfitting.

The Random Forest model obtained a result with an accuracy of 0.790 and ROC AUC of 0.819. Although the ensemble learning approach reduced overfitting from a simple decision tree model, recall and F1-score were lower than Logistic Regression.

XGBoost achieved an accuracy of 0.785 with an ROC-AUC of 0.832, which can be considered competitive with respect to Logistic Regression on the basis of discrimination. Although, Precision, Recall, and F1-Score were slightly lower, which shows that increased complexity in modelling did not bring any better results to the table in terms of classification accuracy.

Discussion

On a whole, Logistic Regression stood out as the model with the highest accuracy, F1-score, and ROC-AUC curve. This shows that the interaction between customer features and churn is mainly linear in nature. Although methods such as Random Forest and XGboost are very effective, in this case, the marginal gain in accuracy did not improve upon the baseline model.

These findings emphasize the fact that simpler models can, at times, outperform more complex models, such as ensembles, if relationships among variables are non-linear in nature. Moreover, the accuracy of Logistic Regression is enhanced by its interpretive capabilities, which make it an ideal candidate for implementation in a telecommunications context.

Table I. Comparison of Model Performance for Gold Deposit Prediction

Comparison of Model Performance for Gold Deposit Prediction					
Model	Accuracy	Precision	Recall	F1-Score	ROC–AUC
Logistic Regression	0.80	0.64	0.57	0.60	0.83
Decision Tree	0.78	0.61	0.52	0.56	0.81
Random Forest	0.79	0.63	0.50	0.56	0.81
XGBoost	0.78	0.61	0.52	0.56	0.83

Chart I. Model Performance Comparison

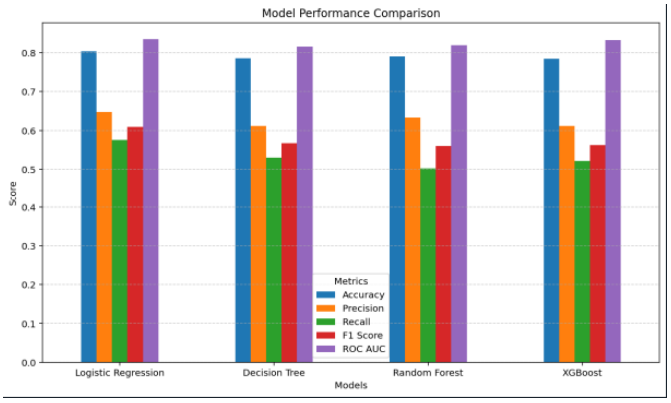
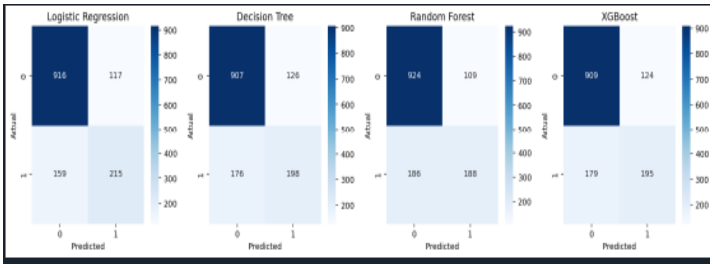


Chart II. Confusion Matrix



VIII. CONCLUSION

In this project, machine learning algorithms were used to model a customer churn model in the telecommunications industry based on demographic information, usage, and billing variables. A whole process, starting from preprocessing, exploration, processing, and evaluation of a model, was performed.

The accuracy of Logistic Regression, Decision Tree, Random Forest, and XGBoost classifier models was contrasted using a variety of evaluation criteria. The output indicated that Logistic Regression performed better than the other sophisticated models despite being less intricate. It can therefore be concluded that customer churn is primarily linear in nature in this given dataset.

The results show that machine learning can be used to aid proactive customer retention and illustrate how model selection can be influenced by data considerations rather than model complexity.

IX. REFERENCES

[1] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proc. ACM SIGKDD, 2016.

[2] I. H. Witten et al., *Data Mining: Practical Machine Learning Tools*, Morgan Kaufmann, 2016.

[3] R. Rodriguez-Galiano et al., “Random Forest for classification,” ISPRS Journal, 2012.

[4] F. Provost and T. Fawcett, “Data Science for Business,” O’Reilly, 2013.

Data Set Link: https://github.com/IBM/telco-customer-churn-on-icp4d/blob/master/data/Telco-Customer-Churn.csv